Proximity data visualization with h-plots

Irene Epifanio[1,*]

1. Departament de Matemàtiques, Universitat Jaume I, Castelló 12071, Spain

* Contact author: epifanio@uji.es

Keywords: $h-plot$, multidimensional scaling, dissimilarity matrix, dimension reduction, human corneal endothelia

Classical multidimensional scaling methods try to preserve all pairwise proximities, whereas many of the recent nonlinear dimension reduction methods, such as Tenenbaum et al. (2000) or Roweis and Saul (2000), use only local neighborhood information to construct a global low-dimensional embedding of a hypothetical manifold near which the data fall (Hastie et al. 2009). Both approaches could become into restrictive constraints in some cases, specially if the measure between objects is not a distance.

Our motivating problem is concerned with the analysis of digital images of human corneal endothelia. In Ayala et al. (2006), different dissimilarities (non-metric measures) between these images were proposed and assessed in a simulation study and, finally, applied to the ophthalmologic problem. Note that triangle inequality is not hold by the dissimilarity considered. In order to compute these dissimilarities, the following libraries of R have been used: *Splancs*; *Spatstat* and *Survival*.

We propose a method based on $h-plot$ (Seber, 1984) for graphical exploration of dissimilarity matrices, which leads to different representations from other methods. It is a non-iterative method, very simple to implement and computationally efficient. The representation goodness can also be easily assessed. It can also be applied to asymmetric proximity data, since our methodology can handle naturally this kind of situation. It has been compared with well known methods and shown its good behavior through several examples, specially with nonmetric dissimilarities. We also propose two alternatives, depending on if the only objective is graphical representation or if cluster and pattern detection is also the goal, using the original dissimilarities or their ranks, respectively.

This work has been done by using free software, R, and specially the library *MASS* (Venables and Ripley, 2002).

An example with more illustrative results on an artificial dataset, the Swiss roll dataset, is available at the following web page: http://www3.uji.es/∼epifanio/RESEARCH/hplot.pdf.

References

J. B. Tenenbaum, V. de Silva, and J. C. Langford (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290 (5500), 2319–2323.

S. T. Roweis and L. K. Saul (2000). Linear embedding nonlinear dimensionality reduction by locally. *Science*, 290 (5500), 2323–2326.

T. Hastie, R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning. Data mining, inference and prediction*. Second Edition. Springer-Verlag.

G. Ayala, I. Epifanio, A. Simó, and V. Zapater (2006). Clustering of spatial point patterns. *Computational Statistics & Data Analysis*, 50(4), 1016–1032.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. ISBN 3-900051-07-0.

B. Rowlingson, P. Diggle, adapted, packaged for R by R. Bivand, pcp functions by G. Petris, and goodness of fit by S. Eglen. *splancs: Spatial and Space-Time Point Pattern Analysis*.

A. Baddeley, and R. Turner (2005). Spatstat: an R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6), 1–42.

T. Therneau, and ported for R by T. Lumley. *survival: Survival analysis, including penalised likelihood.*

G. Seber, (1984). *Multivariate observations*. John Wiley.

W. Venables, and B. Ripley (2002). *Modern applied statistics with S-plus*. Springer.